CLAIMS

What is claimed is:

1. A system that facilitates web-crawling, comprising:

a managing component that performs a predictive analysis in connection with determining if, when, and how to perform web-crawling; and

a web-crawling component that crawls subsets of web pages as a function of the predictive analysis.

2. The system of claim 1, further comprising a decision-theoretic component that makes predictions regarding changes in at least one web page to determine an appropriate time to crawl the at least one web page.

3. The system of claim 2, the decision-theoretic component makes predictions regarding changes in the at least one web page based at least in part on:

a set of possible actions, A, to be performed on the at least one web page;

a set of possible outcomes, O;

a probability that a particular outcome will occur, Pr; and

a utility factor associated with each outcome, Utility(O).

4. The system of claim 3, the decision-theoretic component makes predictions regarding changes in the at least one web page *via* selecting an action, a, from the set of possible actions A, which maximizes the value of:

$$\sum_{o \in O} \Pr(o \mid a) \times Utility(o)$$

where o is an outcome in the set of all possible outcomes, O.

5. The system of claim 1, the predictive analysis is based at least in part on the utility of the at least one web page.

6. The system of claim 1, the predictive analysis is based at least in part on historical data related to the at least one web page.

7. The system of claim 1, the predictive analysis is based at least in part on content contained in the at least one web page.

8. The system of claim 1, further comprising a bundling component that rearranges crawled web pages into new subsets according to the utility of the of the web pages.

9. The system of claim 1, the web-crawling component comprises a Round Robin crawling component that sequentially crawls web pages in a subset and ensures that every web page will be crawled within a crawling period, and a Greedy crawling component that non-sequentially crawls pages according to a score associated with each page.

10. A method for prioritizing web pages for crawling, comprising:

predicting a change in at least one subset of pages;

assigning a score to the at least one subset of web pages;

selectively crawling subsets of pages with higher scores before subsets of pages with lower scores; and

non-selectively crawling all uncrawled pages in sequence.

11. The method of claim 10, further comprising separating at least one subset of pages according to perceived utility of individual pages within the subset.

12. The method of claim 11, further comprising rebundling separated pages into new subsets according to the perceived utility of the separate pages.

13. The method of claim 10, further comprising performing decision-theoretic analysis to determine when to crawl a page based on:

a set of possible actions, A, to be performed on the at least one web page;

a set of possible outcomes, O;

a probability that a particular outcome will occur, Pr; and

a utility factor associated with each outcome, Utility(O).

14. The method of claim 13, further comprising selecting an action, a, from the set of possible actions A, which maximizes the value of:

$$\sum_{o \in O} \Pr(o \mid a) \times Utility(o)$$

where o is an outcome in the set of all possible outcomes, O.

15. The method of claim 10, further comprising making predictions regarding when a web page will change based at least in part on a number of previous crawls of the page.

16. The method of claim 10, the prediction is based on at least one of a universal resource locator (URL) of the page and a web site that points to the URL of the page if the page has not been crawled previously.

17. The method of claim 10, the prediction is based on at least one of a URL of the page, a web site that points to the URL of the page, a hypertext transfer protocol (HTTP) header of the page, and content of the page if the page previously has been crawled only once.

18. The method of claim 10, the prediction is based on at least one of a URL of the page, a web site that points to the URL of the page, an HTTP header of the page, content of the page, history of changes to the page, and delta information related to the page if the page previously has been crawled more than once.

19. The method of claim 18, delta information comprises analyzing the context of the page to determine differences in the page from one crawl to the next.

20. The method of claim 10, the sequence in which all uncrawled pages are non-selectively crawled is based on time since last crawl.

21. The method of claim 20, further comprising ensuring that no web page goes uncrawled for a more than a predetermined time period.

22. The method of claim 10, the score is at least one of a predictive score, a utility score, and a decision theoretic score.

23. The method of claim 10, further comprising weighting a score of at least one web page based on at least one of probability of having changed, maximum average utility, and maximum expected utility.

24. The method of claim 10, further comprising ensuring that no web page is more than D days out of date, where D is a real number.

25. A method for predicting change in a web page via feedback loops, comprising:

selecting a sample set of URLs from subsets of web pages on a server; and

crawling the sample set at regular intervals.

26. The method of claim 25, further comprising employing data gleaned from the crawled sample set to provide training data for learning probability predictors and/or for tuning crawling strategies.

27. The method of claim 25, further comprising employing data gleaned from the crawled sample set to test crawling strategies and/or to build metrics for testing crawling strategies.

28. The method of claim 25, the sample set of URLs is selected from result sets of URLs sent to users utilizing a search engine.

29. The method of claim 28, further comprising weighting more heavily a URL in the result set that has been clicked on by a user than a URL that has not been clicked on by a user.

30. The method of claim 25, further comprising performing a regular crawl on the subsets of web pages prior to selecting the sample set of URLs.

31. The method of claim 30, further comprising recording initial conditions of the sample URLs as determined during the regular crawl.

32. The method of claim 25, further comprising periodically selecting a new sample set of URLs.

33. The method of claim 25, further comprising periodically updating the sample set of URLs by replacing at least one sample URL at a time to gradually create a new sample set of URLs.

34. A method of predicting web page change, comprising:

means for predicting a change in at least one subset of web pages on a web server;

means for crawling the entire subset of pages within a time period;

means for determining a score associated with each page; and

means for selectively crawling pages determined to have a higher score.

35. The method of claim 34, further comprising means for weighting a score of at least one web page based on at least one of probability of having changed, maximum average utility, and maximum expected utility.

36. The method of claim 34, further comprising means for ensuring that no web page is more than D days out of date, where D is a real number.

37. A computer readable medium that has computer executable instructions stored thereon to:

predict a change in at least one web page in at least one subset of web pages on a server;

assign a score to the at least one web page in the at least one subset; and

selectively crawl the at least one web page if the score assigned thereto is above a predetermined minimum value.